1-2017

Online hacker forum censorship: Would banning the bad guys attract good guys?

Qiu-Hong WANG
Singapore Management University, qiuhongwang@smu.edu.sg

Le-Ting ZHANG
Huazhong University of Science and Technology

Meng-Ke QIAO
National University of Singapore

DOI: https://doi.org/10.24251/HICSS.2017.677

Online Hacker Forum Censorship: Would Banning the Bad Guys Attract Good Guys?

Qiu-Hong Wang
Singapore Management University
qiuhongwang@smu.edu.sg

Le-Ting Zhang
Huazhong University of Science and Technology
rudy_zhang@outlook.com

Meng-Ke Qiao
National University of Singapore
mengke@comp.nus.edu.sg

Abstract

To tackle the ubiquitous cybersecurity threats, a few countries have enacted legislation to criminalize the production, distribution and possession of computer misuse tools. Consequently, online hacker forums, which enable the provision and dissemination of malicious cyber-attack techniques among potential hackers or technology-savvy users, are subject to censorship. This project examines the mixed impacts of online hacker forum censorship on users' contribution to protection discussion through a natural experiment with large-scale content analysis. We find that while the enforcement indeed reduced the discussion on malicious cyber-attacks, the discussion on cybersecurity protection could increase or decrease in different scenarios. The rationale is that while the online hacker forum censorship imposes risk to the discussion of malicious attacks, it also reduces the potential benefit from discussing protection issues. Policy implications are discussed.

1. Introduction

Cyber-attack refers to any offence against the confidentiality, integrity and availability of computer data and systems and can range from installing malware on a computers, intruding into or illegally controlling computer information systems to attempting to destroy the infrastructure of entire nations. Cyber-attacks cost the global economy billions of dollars every year, and are growing concerns for businesses and governments around the world [16,21]. One reason for the flooding of cybersecurity violation events is the low cost to acquire the necessary tools and programs to commit cyber-attacks. For example, the online hacker forums enable the communication among potential hackers or technology-savvy users and provide the free-to-access and rich resources on malicious attack techniques. To tackle the ubiquitous cybersecurity threats, a few countries have enacted legislation to criminalize the production, distribution and possession of computer misuse tools. Table 1 provides a list of such countries. Consequently, online hacker forums with the provision and dissemination of malicious attack techniques, are subject to censorship. Banning malicious attack discussion is supposed to increase the knowledge barrier and to reduce the chance of committing cyber-attacks.

Table 1. Countries with legislation on the production/distribution/possession of computer misuse tools

Country	Legislations on the production /distribution /possession of computer misuse tools
Canada	Criminal code, Article
China	Criminal Law
Latvia	Criminal code, Amended Section 244.
Italy	Penal code, Amended Article 615
Lithuania	Criminal code, Amended Article 198
Qatar	Penal code, Part 3 Article 382
Republic of Moldova	Telecommunication Law, Article 66
Russian Federation	Criminal Code, Act 273 and 138.1
Saint Lucia	Criminal Code, Article 330, 331

While few opponents would rise against the regulation on disseminating bomb making information, the same rationale may not be expected to malicious attack discussion. The ambiguous opinions towards the dissemination of malicious attack techniques are rooted in the distinctions between conventional crimes and cyber-attacks. First, malicious attack discussion plays a dual role in protection and attack [29]. For example, the port scanners and exploit tests are powerful instruments for network administrators to detect their information system vulnerabilities, and at the same time the detected vulnerabilities could be exploited by hackers to commit cyber-attacks. In fact the endless combat between cyber-attacks and its countermeasures becomes the driving force for the advancement of defensive technology. Second, the

URI: http://hdl.handle.net/10125/41840
ISBN: 978-0-9981331-0-2

community, as the perpetrators against cyber security, from its origins in 1960s, were considered the "Heroes of the Computer Revolution" [20]. Through decades of migration, hacker community have become a congregation of the "white hats", "black hats" and "gray hats". The white hats have commitment to information freedom, mistrust of authority, and heightened dedication to meritocracy. The black hats are engaged with forbidden actions including mockery, spectacle, and transgression. The gray hats participate in both black and white domains [10]. Thus it is lack of a clear moral judgment about hackers. Lastly, the loss rendered by cyber-attacks is largely intangible and hard to measure. All of the aforementioned factors contribute to the debate on online hacker forum censorship. In this study we address a straightforward question:

What is the impact of banning malicious attack discussion on users' contribution to protection discussion in online hacker forums?

The answer to this question is not straightforward given the intertwining of contesting and conquering between malicious attack discussion and protection discussion. Banning malicious attack discussion is supposed to increase the knowledge barrier and to reduce the chance of committing cyber-attacks. On the other side, lack of the alert from malicious attack discussion, forum users may become less interested or poorly motivated to attend protection discussion. If banning malicious attack discussion discourages the contribution on protection discussion and thus reduces the public's awareness of potential threats and technical measures against malicious attack, its role in deterring cybersecurity threats may not be well justified. Instead banning malicious attack discussion on online hacker forums may force the black hat back to the underground hacker communities, thus making the potential cybersecurity threats invisible to the public and hence hard to be tackled.

We investigate the research question in the context of the Chinese online hacker forums. On Feb 28,2009, China government enacted the Amended Article 285 in the Criminal Law which states that "Whoever provides programs or tools specially used for intruding into or illegally controlling computer information systems, or whoever knows that any other person is committing the criminal act of intruding into or illegally controlling a computer information system and still provides programs or tools for such a person shall, if the circumstance is serious, be punished under the preceding paragraph".[1] Following the enforcement of this amendment, the Internet security agencies in China conducted intensive censorship to online hacker forums. Forum administrators also removed considerable amount of posts containing malicious techniques and regulated the forums with strict rules and surveillance on user-generated contents. As a result, the number of posts on malicious attack in each of our studied two forums has significantly dropped from then onwards. We examine the change of the number of posts on protection before and after the enforcement of the Amended Article 285 at the forum aggregate level and the user group level. Innovative text mining and content classification techniques have been applied into the data processing.

We find that while the enforcement indeed reduced the discussion on malicious cyber-attacks, the discussion on cybersecurity protection could increase or decrease in different scenarios. The rationale is that while banning discussion the online hacker forum censorship imposes risk to the discussion of malicious attacks, it also reduces the potential benefit from discussing protection issues.

This paper is organized as follow. Section 2 is about the related literatures. Section 3 introduces the context of this study. In section 4, we describe our classification method. Section 5 reports the empirical analysis and estimation results. Section 6 concludes the study with discussion about implication and limitation.

2. Related literature

This study is related to three streams of research in the literature including hacker behavior, Internet regulation, and hacker forum text analysis.

2.1. Hacker behavior

Hackers can be classified as white hats or black hats based partly on their intents and the potential criminal nature of their activities. Individuals who attempt to hack into computer systems and ruin the systems are referred to as black hat hackers; individuals who attempt to protect the computer systems are known as ethical hackers or white hat hackers [27].The earliest white hats can be traced back to the late 1960s with the belief that computers can be the basis for beauty and a better world [20]. Following the growth of white hats, black hats evolved from the telephone phreakers to the computer hackers [10]. However, the white hats and black hats are not so distinct from each other. White hat hackers could simulate the attacks used by black hat hackers in order to test potential security risks and understand how to defend against them [9]. Black hat hackers can be

1 http://www.gov.cn/flfg/2009-02/28/content_1246438.htm

recruited to develop security software or to provide IT security consultancy service [4]. And there exist the gray hat hackers who lie between the white and black hats, committing to security by hacking into the political territory [10]. Hence, the moral judgment about hacker is ambiguous.

Hacker's moral ambiguity is consistent with their communications in online hacker forums. The participants in hacker forums discuss issues about both malicious attack and protection. They may post step-by-step guide to help others conduct malicious attacks, e.g. SQL injection, web exploits, and decryption [6]. Exploit tools or malwares are also available as attachments, e.g. the Dirt Jumper DDos attack, keyloggers and crypters [25]. They also discuss technologies, methodologies and practices about detecting, preventing and tracking the black hats to protect information assets.

Being aware of the moral ambiguity among hackers, to the best of our knowledge, no previous work has addressed the interdependency between white hats and black hats.

2.2. Internet regulation

A number of countries have enacted policies to regulate the Internet which enables the generation, communication and dissemination of both benign and malicious content. They block access to the Internet content and websites which are harmful to the public [2]. For instance, the contents about hate speech are restricted by the France government [5]. Websites threatening national security are blocked in South Korea and Pakistan [11]. The creation of hacking tools is considered a criminal offense in the United Kingdom and Germany. On Feb 28, 2009, China has enacted the Amended Article 285 of its Criminal Code which criminalizes the provision of hacking tools or programs. The neutrality pertaining to information technology leads to the debate on regulation. For example, encryption has the potential to further massive terrorism and facilitate greater security in communication. Thus some of the law enforcement communities advocate its criminalization but others stand by accessing to the technology [18]. In our case, hackers are two-sided, playing positive and negative roles in cybersecurity, and sharing both malicious attack and protection knowledge. Due to law enforcement, some black hats may quit from the censored online hacker forums. As a result, forum users may become less interested in contribution simply due to the shrinking group size [30]. And lack of the alert from malicious attack discussion, forum users may become less interested or poorly motivated to attend protection discussion. It's unclear whether forbidding malicious attack discussion forfeits their contribution to protection discussion [18]. Hence, it is important to figure out what impact banning malicious attack discussion could have on the contribution to protection discussion.

2.3. Hacker forum text analysis

Different from the underground hacker communication channel, i.e., ICQ, where the observations are limited by personal contacts, hacker forums are the publicly accessible hacker communities where the vast amount of user-generated content can be investigated in a longitudinal base. However, unlike online product review where the user-generated content is structured or semi-structured, the unstructured and diversified contents in hacker forums impose great challenge to quantitative analysis. Most of the relevant text analysis studies are focused on uncovering the dark side of the mysterious group. Abbasi et al. [1] use an interaction coherence analysis (ICA) framework to identify expert hackers in forums. Samtani et al. [25] apply classification and topic modeling techniques to investigate the functions and characteristics of assets in hacker community. In order to have better understanding of hacker terms and concepts, Benjamin et al. [8] utilize recurrent neural network language models (RNMLM) to model language. To the best of our knowledge, no previous work has distinguished the hacker forum posts by hackers' intents of either malicious attacks or protection. Thus posts on protection are mostly ignored.

In this study, we classify posts into three categories, malicious-attack, protection, or the irrelevant through supervised machine learning. With human-labeled training datasets, we use n-gram, weight, together with information gain [26, 24] to generate and select features, then feed them into Naive Bayes and SVM classifiers. We choose Naive Bayes and SVM as the classifiers because they are classical and can be adopted in many occasions. SVM also often reported best performance in many previous online text classifications [31]. At last, classifiers with good precision and recall rate are used to label the remaining posts.

3. Context and Theory Discussion

3.1. Hacker forums

With the consideration on popularity, established period, the theme of the forum and major topics, we choose forum A and forum B among the most representative hacker forums in China as the research

subjects, and investigate the impact of banning malicious attack discussion on participants' contribution to protection discussion. According to the web traffic ranking by Alexa.com, Forum A and Forum B are ranked the second and third respectively in the Chinese hacking category. [2] The No.1 forum, established in 2008, cannot provide a balanced longitudinal dataset with enough time periods before and after the enforcement of the Amended Article 285. Forum A was established in March 2001, one of the earliest and most famous hacker forums in China. It aims to cultivate hackers with advanced knowledge and techniques and hence has long enjoyed a great popularity. Different from forum A, Forum B, established in December 2002, aims to raise people's awareness of cyber security and to provide related services. Posts on either malicious attacks or protection are found in both forums, perhaps due to the ambiguous roles of hackers. But the different value propositions have resulted in more discussion on malicious attacks in Forum A and more discussion on protection in Forum B.

3.2. The Amended Article 285 in the Criminal Law of People's Republic of China

On Feb 28,2009, Chinese government enacted the Amended Article 285 in the Criminal, which states that "Whoever provides programs or tools specially used for intruding into or illegally controlling computer information systems, or whoever knows that any other person is committing the criminal act of intruding into or illegally controlling a computer information system and still provides programs or tools to such a person shall, if the circumstances are serious, be punished under the preceding paragraph". The enforcement of this amendment has generated widespread and substantial impacts on the online hacker forums in China. First, the Internet security agencies in China conducted intensive censorship to online hacker forums. The chief administrator of forum A was even arrested and sentenced to five-year prison. Second, to comply with this law, many hacker forum administrators implemented a series of regulations to forum participants, including deleting posts on malicious attack, promulgating more rigorous content censorship and alerting those participants who disseminated malicious attack discussion and tools in online forums. Given the dual usage of hacking techniques and the ambiguous incentives of hackers, it is not clear how the law enforcement against malicious attack discussion will indirectly affect the participants' contribution to protection discussion.

3.3. Theory Discussion

We use the volume and ratio of posts to measure forum users' contribution on discussion, as the ratio can offset any change in the overall contributions across the whole forum. Our hypotheses are based on three main effects resulted from the law enforcement.

Displacement effect. Displacement effect in this study means that forum users who would have attended discussion on malicious attack may instead choose to discuss protection issues. This is related to the communication and technical interests pertaining to the participants in the hacker forums. First, meritocracy is emphasized in their active area [12,13] and hackers acquire reputation which accumulated from their activity levels and post quality [7]. For successful hackers, they do feel the need to brag and share their accumulated knowledge [12, 17]. Second, hackers are technology savvy while both hacking and protection share the same technical foundation. Considering the risk of discussing malicious attacks, they may convert discussing hacking knowledge into discussing protection knowledge, and continue to launch posts on protection, in order to keep active and accumulate their reputation in forums. As a result, banning malicious attack discussion may lead to more posts on protection.

New user effect. As the hacker forums become more protection oriented, it would attract new users who are interested in protection techniques. As a result, there would be more white hats in hacker forums than before. Thus the amount of posts on protection and the ratio of posts on protection to all posts would increase.

Both displacement effect and new user effect support the positive effect of banning malicious attacks discussion on the contribution of protection discussion. Thus we expect the number of posts on protection increases after the law enforcement and the extent of increase is larger than the other irrelevant posts in the forum.

Precaution reduction effect. Posts on malicious attack may raise the precaution awareness and stimulate the discussion on protection issue. The law enforcement deters forum users from discussing malicious attack, and a large number of posts on attack were deleted by forum administrators. This may reduce the attention and interests on protection issues. Therefore, the volume and ratio of posts on protection may decrease.

Hence, what impact could banning malicious discussion and tools have on posts on protection is a pending question subject to empirical test.

[2] In Alexa.com, hacking is listed as one of the sub-categories in Computers. Ranking was assessed on April 5, 2016

4. Data processing

4.1. Definition of intents

For the purpose of our research, we classify the intents of posts into three categories. The first is "malicious attack", which means the post contains malicious attack intent, expressing a tendency to attack others; the second is "protection", which is about measures of protecting personal or company (information, account) from being attacked by malicious hackers; the third is "irrelevant", for those neither related to "malicious attack" nor to "protection". Through a thorough study on hacker forum posts, we summarize the typical topics of each category in Table 2. After defining the specific contents in each category, text classification is needed to label each post accordingly.

Table 2. Typical Topics and Post Examples

Malicious attacks
Typical Topics
footprinting and reconnaissance, scanning networks, enumeration, system hacking, Trojans and back- doors, viruses and worms, sniffers, social engineering, denial of service, session hijacking, hacking web servers, hacking web applications, SQL injection, hacking wireless networks, evading IDS, IPS, firewalls, and honeypots, buffer overflow, and cryptography
Post Examples
Postid=52972, "Recently, I scanned out a ROOT blank command of a host MSSQL, how can I get the host's administrator right" Postid=3045218, "Numerous ways to surf internet for free in internet bar"!!!!!
Protection
Typical Topic
How to defense from hackers' attacks, including installation and setting of firewall, closing certain ports
Post Examples
Postid=2754943, "Help….My computer has been infected by virus." Postid=3228449, "Share: How to protect IP from being stolen"
Irrelevant
Typical Topic
Other contents that are not relevant to attack or defense. For example, basic computer operation, chatting, advertisement
Post Examples
Postid=26837, "How to run DOS under windows 2000 " Postid=2808442, "Good news! Tencent is celebrating 6th anniversary now, 6 digit QQ number can be applied for free. Apply for it soon!"

4.2. Text classification

The whole text classification process is presented in Figure 1. Since a leading post represents the topic of a whole thread, we constrained our samples to all of the leading posts in the two forums. Two human annotators, also as the co-authors of this study, independently labeled 18833 leading posts out of the 140802 leading posts in Forum A and 5459 leading posts out of the 28317 leading posts in Forum B. Both of them including one postgraduate and one senior undergraduate, are majored in information systems, and have received more than six-month training on the domain knowledge of information security and hacker communities before working on labeling. Their inter-rater agreement, using kappa statistics, is 0.778 for Forum A and 0.92 for Forum B, which suggests sufficient inter-rater reliability. We then use the labeled dataset as the training dataset and testing dataset.

The next step is to preprocess these unstructured texts. Unlike English, Chinese does not have space between words. So we first need to segment each sentence into tokens via Rwordseg provided in R. Meanwhile, stop words, useless in this classification task, are removed. We then use N-grams to generate more features. To select features, we give higher weights on post title and use information gain to filter out less important features while reserving those that are more useful in discriminating posts [15, 19]. Then these feature sets are used to train Naive Bayes and SVM classifiers. Following classifier training, we use 10-fold cross validation to evaluate the performance of the classification. Finally, for each sub forum, classifiers with the best performance are applied to labelling the remaining posts.

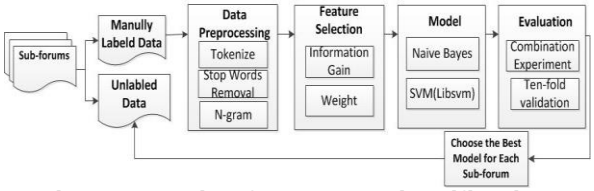

Figure1. Hacker forum text classification process

The classification is implemented by Rapidminer with performance reported. For Forum A, the average precision, recall and F1-measure of three classes are 86.36%, 80.11% and 82.73% respectively; For forum B, the average precision, recall and F1-measure of three classes are 77.83%, 71.23% and 74.24% respectively. Since no previous study has classified the intents of posts in hacker forums, no existing benchmark could be applied. Referring to a recent study which identified users' intents in online health forum using word vector and SVM in text classification [28], their average precision, recall and F1-measure of all classes are 49.77%, 48.44% and 48.78% respectively.

5. Model and empirical analysis

5.1. Model and description

We address our research question at both aggregate level and user level. Model 1 at the aggregate level

investigates how the daily volume (ratio) of posts on protection (PoP) changes with the law enforcement (the banning of malicious attack discussion).

$$PoP_t = \alpha_0 + \alpha_1 E_t + +ContralVars_t + TimeTrend_t + Lag_t + \varepsilon_t \quad (1)$$

where *t* denotes date *t*, PoP_t is the daily amount of PoP in a forum , E_t indicates the enforcement of the amended Article 285. $ControlVars_t$ is a vector consisting of the daily number of post users and the daily number of new users, to control the impact of forum group size on post contribution [30]. $TimeTrend_t$ captures the time trend. Lag_t is the first order lag of the dependent variable. Excluding ratios, all variables are converted to the logarithmic form.

The Heckman model is employed to analyze the impact of the law enforcement on the ratio of protection posts. We calculate the ratio as the amount of PoP over the total amount of PoP and irrelevant posts. The malicious attack posts are excluded from the denominator as they have been seriously manipulated following the law enforcement. Further the first stage of the Heckman model can capture the impact of the law enforcement on the probability of posting or not posting. In order to correct the selection bias due to no leading post in a forum at some days, we calculate the inverse Mills ratio based on the estimation result in the first step, and incorporate it into stage 2. The Heckman model is specified as following,

$$\textbf{Stage1}: IsPost_t = \beta_0 + \beta_1 E_t + ControlVars_t + TimeTrend_t + Lag_t + \mu_t \quad (2)$$

$$\textbf{Stage2}: Ratio\_Pop_t = \gamma_0 + \gamma_1 E_t + ControlVars_t + TimeTrend_t + LagS_t + \tau_t \quad (3)$$

In equation (2), $IsPost_t$ is a dummy variable that equals to 1 if at least one post was posted at day *t*, and 0 otherwise. Lag_t is the first order lag of the total amount of posts. In equation (3), $LagS_t$ is a vector of the first order lag of the daily amount of PoP and the total amount of posts. Other variables have the same meanings as in equation (1).

Model 2 at user level investigates the change of the users' contribution to protection posts before and after the law enforcement. We constrain the subjects to users who joined the forum before Feb 28, 2009 and assort users with the same joining date into one group.[3] To ensure the symmetric time window before and after the enforcement date for each group, we drop groups who joined the forum before 2005. We finally get 1842 groups for Forum A and 1217 groups for Forum B. For each group, their time windows before and after the law enforcement equal to the number of days between their joining date and the enforcement date. For example, for a group of users who joined the forum in Jan 1, 2009, the number of days before the law enforcement is 58 days. Thus we only check their contributions within 58 days after the law enforcement. We check how the number of PoP by group *i* change before and after the enforcement date using a fixed-effect model,

$$GroupPoP_{it} = \alpha_0 + \alpha_1 E_t + ControlVars_{it} + \epsilon_{it} \quad (4)$$

where *t* equals to 0 for the time window before the enforcement date and 1 otherwise. We use the other groups' total amount of posts/replies on protection and total amount of irrelevant posts/replies to control for any impact due to the forum size and peer influence. Same as model 1, when the dependent variable is the ratio of PoP, the Heckman model is applied.

$$\textbf{Stage1}: IsPost_{it} = \beta_0 + \beta_1 E_t + ControlVars_{it} + \mu_{it} \quad (5)$$

$$\textbf{Stage2}: RatioGroupPoP_{it} = \gamma_0 + \gamma_1 E_t + ControlVars_{it} + \tau_{it} \quad (6)$$

We derive the inverse Mills ratio from stage 1 and incorporate it into stage 2. In stage 2, besides those control variables in stage 1, vector $ControlVars_{it}$ also includes the amount of group users and the length of time window to control for the effects due to group size and time interval. Tables 3.1 and 3.2 report summary statistics for main variables used in model 1 and model 2 respectively.

5.2. Forum aggregate level analysis

The columns 1-3 and 4-6 in Table 4 report the regression results of model1 for forum A and forum B respectively. In Column 1 and Column 4, the coefficient of the law enforcement for Forum A is positive and significant while it is negative and significant for Forum B. These results seem conflicting with each other but are reasonable given the different positioning of Forum A and Forum B. As introduced in Section 3.1, Forum A aims to cultivate hackers with advanced knowledge and techniques while Forum B aims to raise people's awareness of cyber security and provide related services. Hence banning the malicious discussion increases the perceived risk for the black hats in Forum A but at the same time reduces the perceived benefit for the white hats in Forum B. Consequently, the displacement effect explains the positive and significant coefficient of the enforcement indicator for Forum A while the precaution reduction effect explains the negative and significant coefficient of the enforcement indicator for Forum B.

[3] We group users by joining dates because the size of individual level data is too big. There are 159626 unique users in forum A and 37307 unique users in forum B. In our next stage of this research, we will conduct individual-level analysis.

For Heckman model, both of the results in Column 3 and Column 6 show that the ratio of the PoP increases significantly after the law enforcement. This suggests that banning malicious attack discussion generates relatively more positive effect on protection discussion than discussion on issues irrelevant with attack and protection. However, referring to Column 2, the coefficient of the enforcement indicator is negative and significant while it is negative and insignificant in Column 5. This difference further suggests the distinct responses of users in Forum A compared to those in Forum B. Combining the results in Columns 1,2,4 and 5, it shows after the law enforcement, users in Forum A which consisted of more black hats relative to White hats, choose to either keep mute or discuss protection issues. Differently, in Forum B which consisted of more white hats relative to black hats, users may keep posting but the total number of PoP reduced.

5.3. User group level analysis

By splitting users into groups based on joining date, we are able to examine the change in amount and ratio of PoP at group level, in particular for old users who joined the forums before the enforcement. Table 5 reports the regression results of model 2. Generally, the results of model 2 are consistent with that of model 1 presented in Table 4, i.e. the coefficients of the enforcement indicator in columns 1-5 of Table 5 are significant, with the same sign as the corresponding specifications reported in columns 1-4 of Table 4. The main difference is that the negative effects of banning malicious attack discussion on general discussion, in particular for discussion on protection, becomes more salient in model 2, e.g., columns 2,6,7 and 8 in Table 5. These evidences together with results in columns 2, 5 and 6 of Table 4 further clarify that the increasing ratio of PoP as reported in column (6) of model 1 is mainly due to the contribution from new users who joined Forum B after the enforcement.

To explain the distinct results from data in Forum A and Forum B, we conduct a paired t-test to compare the daily number of PoP posted by old and new users in Forum A and Forum B respectively. Table 6 shows that on average, users in Forum B contribute more PoP than users in Forum A, which suggests the systematic difference of user profiles in the two forums. Further, the mean number of PoP posted by new users in Forum B is much more than that of PoP posted by old users, which further suggests that the positioning of Forum B effectively attracts more white hats than forum A

6. Conclusion and implications

Combining the statistics in Table 6 with the regression outcomes, we can conclude that while banning malicious attack discussion imposes risk to the discussion of malicious attacks, it also reduces the potential benefit from discussing protection. Thus the black-hat hackers may respond to the enforcement by switching to discussing protection topics; while the white-hat hackers become less motivated to discuss protection issues. As a result, the impact of online hacker forum censorship is a mix which depends on user profile in each forum.

Internet censorship is a very important and sensitive issue to policy makers. This study shows that the bad guy and good guy may not be always substitutes to each other. Instead they are interdependent and their boundaries may become ambiguous due to technology neutrality and the ethical ambiguity pertaining to hacker community. In particular, we find that banning malicious attack discussion discourages the contribution on protection discussion by the white hats. On the other side, to reduce the probability of being punished, the black hats may approach the underground hacker communities for discussing malicious attacks. Thus the potential cybersecurity risk imposed by malicious cyber-attack discussion does not really reduce but just becomes less observable. This is an even worse situation since the public become less alerted about the potential threats and are also less aware about the technical countermeasures against malicious attacks. Hence, the role of the online hack forum censorship in deterring cybersecurity threats may not be well justified. Instead of banning malicious attack discussion in online hacker forums, our study proposes that the authorities should encourage more discussion about the disclosure of cyber-attack threats and their countermeasures. Banning the bad guys does not attract and help good guys, but pushes the devil to the dark.

This study can be improved through at least three ways. The first is to improve the performance of text classification through in-depth machine learning. The second is to broaden the coverage of hacker forums in order to capture any interdependence and enhance its generalizability. Lastly the current empirical model should be improved by including individual level analysis.

Acknowledgement This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Grant Approval No. 16-C220-SMU-002).

7. References

[1] A. Abbasi, W. Li, V. A. Benjamin, S. Hu, and H. Chen, "Descriptive Analytics: Examining Expert Hackers in Web Forums", JISIC, 2014, pp. 56-63.
[2]Y. Akdeniz, "To block or not to block: European approaches to content regulation, and implications for freedom of expression", Computer Law & Security Review, 26 (2010), pp. 260-272.
[3] N. Archak, A. Ghose, and P. G. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews", Management Science, 57 (2011), pp. 1485-1509.
[4] N. Auray, and D. Kaminsky, "The professionalisation paths of hackers in IT security: The sociology of a divided identity", Annales Des T d écommunications, Springer, 2007, pp. 1312-1326.
[5] D. E. Bambauer, "Filtering in Oz: Australia's foray into Internet censorship", U. Pa. J. Int'l L., 31 (2009), pp. 493.
[6] V. Benjamin, and H. Chen, "Developing understanding of hacker language through the use of lexical semantics", Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on, IEEE, 2015, pp. 79-84.
[7] V. Benjamin, and H. Chen, "Securing cyberspace: Identifying key actors in hacker communities", Intelligence and Security Informatics (ISI) , 2012 IEEE International Conference on, IEEE, 2012, pp. 24-29.
[8] V. Benjamin, W. Li, T. Holt, and H. Chen, "Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops", Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on, IEEE, 2015, pp. 85-90.
[9] T. Caldwell, "Ethical hackers: putting on the white hat", Network Security, 2011 (2011), pp. 10-13.
[10] E. G. Coleman, "Hacker", In M. L. Ryan, L. Emerson, and B. J. Robertson (Eds.)， The Johns Hopkins Guide to Digital Media. Johns Hopkins University Press, 2014, pp. 245-249
[11] R. Faris， and N. Villeneuve, "Measuring global Internet filtering", Access denied: The practice and policy of global Internet filtering, 5 (2008).
[12] T. J. Holt, "Subcultural evolution? Examining the influence of on-and off-line experiences on deviant subcultures", Deviant Behavior, 28 (2007), pp. 171-198.
[13] T. J. Holt, D. Strumsky, O. Smirnova, and M. Kilger, "Examining the social networks of malware writers and hackers", International Journal of Cyber Criminology, 6 (2012), pp. 891.
[14] L. Hong, and B. D. Davison, "A classification-based approach to question answering in discussion boards", Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2009, pp. 171-178.
[15] P. J.-H. Hu, T.-H. Cheng, C.-P. Wei, C.-H. Yu, A. L. Chan, and H.-Y. Wang, "Managing clinical use of high-alert drugs: A supervised learning approach to pharmacokinetic data analysis", IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37 (2007), pp. 481-492.
[16] P. Hyman, "Cybercrime: it's serious, but exactly how serious?", Communications of the ACM, 56 (2013), pp. 18-20.
[17] T. Jordan, and P. Taylor, "A sociology of hackers", The Sociological Review, 46 (1998), pp. 757-780.
[18] N. K. Katyal, "Criminal law in cyberspace", University of Pennsylvania Law Review, 149 (2001), pp. 1003-1114.
[19] M. Koppel, and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution", Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 2003, pp. 72.
[20] S. Levy, Hackers: Heroes of the computer revolution, Penguin Books, New York, 2001.
[21] McAfee, Net Losses: Estimating the Global Cost of Cybercrime, Economic Impact of Cybercrime II, Center for Strategic and International Studies 2014. http://mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf
[22] B. Pang, and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts ", Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, pp. 271.
[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79-86.
[24] J. R. Quinlan, "Induction of decision trees", Machine learning, 1 (1986), pp. 81-106.
[25] S. Samtani, R. Chinn, and H. Chen, "Exploring hacker assets in underground forums ", Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on, IEEE, 2015, pp. 31-36.
[26] C. E. Shannon, "A mathematical theory of communication", ACM SIGMOBILE Mobile Computing and Communications Review, 5 (2001), pp. 3-55.
[27] G. Steube, "A logistic regression model to distinguish white hat and black hat hackers".Diss.Capella University,2004.
[28] T. Zhang, J. H. Cho, and C. Zhai, "Understanding user intents in online health forums", Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM, 2014, pp. 220-229.
[29] X. Zhang, A. Tsang, W. T. Yue, and M. Chau, "The classification of hackers by knowledge exchange behaviors", Information Systems Frontiers, 17 (2015), pp. 1239-1251.
[30] X. Zhang, and F. Zhu, "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia", The American economic review, 101 (2011), pp. 1601-1615.
[31] Y. Zhang, Y. Dang, and H. Chen, "Gender classification for web forums", IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 41 (2011), pp. 668-677.

s

Table 3.1 Descriptive Statistics of Model 1

Forum	Variable	Pre-enforcement				Post-enforcement			
		Mean	Std.dev.	Min	Max	Mean	Std. Dev.	Min	Max
A	PoP	3.651	3.428	0	20	0.983	1.405	0	16
	Ratio of PoP	0.040	0.028	0	0.194	0.032	0.049	0	0.4
	No. of post users	369.171	121.530	0	965	293.513	185.182	0	1304
	No. of new users	58.895	68.983	0	1222	31.058	74.453	0	2075
	Total no. of PoP and irrelevant posts	86.554	46.874	0	259	28.728	20.770	0	123
	No. of days	2191							
B	PoP	4.517	6.983	0	67	5.044	5.384	0	50
	Ratio of PoP	0.444	0.322	0	1	0.276	0.181	0	1
	No. of post users	35.144	43.913	0	173	86.625	41.985	0	289
	No. of new users	7.326	13.183	0	250	13.914	20.062	0	274
	Total no. of PoP and irrelevant posts	13.128	19.099	0	132	17.448	13.457	0	119
	No. of days	1900							

Table 3.2 Descriptive Statistics of Model 2

Forum	Variable	Pre-enforcement				Post-enforcement			
		Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
A	No. of PoP by group i	3.318	5.827	0.000	148.000	0.194	1.299	0.000	32.000
	No. of days	928.838	536.729	1.000	2067.000	928.838	536.729	1.000	2067.000
	No. of new joined users	51382.390	42310.450	52.000	117386.000	28065.040	9504.100	63.000	42209.000
	No. of group users	63.726	63.489	1.000	1222.000	63.726	63.489	1.000	1222.000
	No. of irrelevant posts by other groups	65374.250	41430.920	48.000	118346.000	23910.240	8325.205	52.000	31491.000
	No. of PoP by other groups	2998.267	2242.678	1.000	6117.000	850.834	275.272	1.000	1093.000
	No. of irrelevant replies by other groups	545692.700	277299.500	604.000	863592.000	391994.900	140379.300	594.000	5257020
	No. of replies on protection by other groups	17938.680	12036.800	8.000	34119.000	6786.086	2217.158	27.000	8521.000
	No. of groups	1842							
B	No. of PoP by group i	4.198	22.303	0.000	352.000	0.472	6.299	0.000	201.000
	No. of days	680.933	393.632	1.000	1612.000	680.933	393.632	1.000	1612.000
	No. of new joined users	5716.440	2323.887	5.000	8328.000	10467.150	6682.378	4.000	27024.000
	No. of group users	6.837	12.807	1.000	250.000	6.837	12.807	1.000	250.000
	No. of irrelevant posts by other groups	8297.601	2683.673	2.000	10075.000	9239.056	6110.914	17.000	23187.000
	No. of PoP by other groups	3995.307	1453.264	9.000	5285.000	3577.674	2162.968	10.000	7568.000
	No. of irrelevant replies by other groups	49332.82	15094.300	56.000	58218.000	96695.570	46176.240	87.000	170131.000
	No. of replies on protection by other groups	13505.490	4494.019	46.000	17269.000	20892.090	10265.840	43.000	36383.000
	No. of groups	1217							

Table 4. The Estimation Result of Model 1

	(1)	(2)	(3)	(4)	(5)	(6)
	Forum A	Forum A Stage 1 of Heckman	Forum A Stage 2 of Heckman	Forum B	Forum B Stage 1 of Heckman	Forum B Stage 2 of Heckman
VARIABLES	No. PoP	Is there any post?	Ratio of PoP	No. PoP	Is there any post?	Ratio of PoP

The Enforcement indicator	0.239*** (0.058)	-1.056*** (0.288)	0.038*** (0.009)	-0.426*** (0.062)	-0.460 (0.371)	0.051** (0.024)
No. of new users	0.0274* (0.017)	-0.115*** (0.032)	2.49e-06 (0.002)	-0.0236 (0.017)	-0.352*** (0.103)	0.000755 (0.007)
No. of post users	0.107*** (0.022)	1.017*** (0.092)	0.030*** (0.004)	0.399*** (0.020)	1.241*** (0.107)	-0.059*** (0.011)
No. of PoP at day t-1	0.215*** (0.021)		0.0250*** (0.003)	0.303*** (0.021)		0.097*** (0.013)
Total no. of posts on protection or irrelevant posts at day t-1		-0.655*** (0.005)	-0.060*** (0.004)		0.193*** (0.058)	-0.079*** (0.014)
Linear time trend	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Constant	0.707*** (0.133)	0.122 (0.405)	0.151*** (0.022)	-0.301*** (0.046)	-1.634*** (0.163)	0.739*** (0.026)
Observations	2,191	2,191	2,191	1,900	1,900	1,900
Adj. R-squared	0.531	0.171		0.519	0.239	

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 5. The Estimation Result of Model 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Forum A	Forum A Stage 1 of Heckman	Forum A Stage 2 of Heckman	Forum A Stage 2 of Heckman	Forum B	Forum B Stage 1 of Heckman	Forum B Stage 2 of Heckman	Forum B Stage 2 of Heckman
VARIABLES	No. PoP	Is there any post?	Ratio of PoP	Ratio of PoP	No. PoP	Is there any post?	Ratio of PoP	Ratio of PoP
	Fixed -effect		Fixed -effect	OLS	Fixed -effect		Fixed -effect	OLS
Enforce law indicator	1.556*** (0.121)	-1.339*** (0.391)	0.059*** (0.025)	0.060*** (0.019)	-0.925*** (0.143)	-2.341** (0.332)	-0.408*** (0.153)	-0.365*** (0.110)
No. of users in group i		-1.120*** (0.212)		-0.007 (0.004)		1.039*** (0.075)		0.049* (0.030)
No. of days		-1.661*** (0.190)		0.018 (0.012)		0.287 (0.222)		0.308*** (0.062)
No. of PoP by other groups	-2.976*** (0.359)	0.0172 (1.051)	-0.188** (0.082)	-0.102* (0.052)	-0.912*** (0.368)	0.423 (0.727)	-0.0729 (0.308)	-0.157 (0.221)
No. of replies on protection by other groups	-1.978*** (0.271)	-1.213 (0.946)	0.286*** (0.0669	0.138*** (0.0462)	-0.548 (0.520)	-1.985** (1.011)	-0.381 (0.398)	0.0739 (0.270)
No. of irrelevant posts by other groups	9.449*** (0.460)	3.526** (1.533)	0.0837** (0.116)	0.106 (0.072)	0.783*** (0.262)	-0.962** (0.489)	-0.274 (0.227)	-0.217 (0.175)
No. of irrelevant replies by other groups	-2.855*** (0.334)	-0.651 (0.765)	-0.269** (0.080)	-0.153*** (0.044)	1.032** (0.495)	3.100*** (0.796)	0.950** (0.379)	0.192 (0.228)
the inverse Mills ratio			-0.003** (0.002)	-0.002* (0.001)			0.020 (0.029)	-0.029 (0.020)
Constant	-17.0***	-8.596***	1.238*	0.377***	-1.705**	-5.557***	-0.563	-0.405
	(2.537)	(2.369)	(0.657)	(0.116)	(0.966)	(1.483)	(1.152)	(0.501)
Observations	3,684	3,684	2,960	2,960	2,434	2,434	897	897
R-squared	0.695		0.046		0.265		0.080	
Number of groups	1,842	1,842	1,771	1,771	1,217	1,217	709	709

Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6. Summary Statistics for PoP

Forum	Old Users		New Users	
	Mean	Standard Error	Mean	Standard Error
A	0.473	0.028	0.511	0.030
B	1.254	0.067	3.428	0.117